Determining light verb constructions in contemporary British and Irish English

Ronan, Patricia ; Schneider, Gerold

Abstract: This study implements an automated parser-based approach to the investigation of light verb constructions. The database consisting of ICE-GB and ICE-IRE is used to obtain qualitative and quantitative results on the use of light verb structures. The study explains and evaluates the steps employed to optimize parser output in detecting open lists of light verb constructions. It discusses the qualitative usage differences of these structures in the data between the two varieties and finds that ICE-GB favours fewer high frequency light verbs while ICE-IRE contains more diverse lower frequency light verbs and more passives. Overall, counts of light verb constructions are considerably higher than previously assumed. The projected counts suggest that attestations of light verb constructions will increase considerably if the search is not restricted to certain high-frequency light verbs as is typically done in studies employing manual or semi-automatic approaches to data collection.

DOI: https://doi.org/10.1075/ijcl.20.3.03ron

Determining Light Verb Constructions in Contemporary British and Irish English

Abstract

This study implements an automated parser-based approach to the investigation of light verb constructions. The data base consisting of ICE-GB and ICE-IRE is used to obtain qualitative and quantitative results on the use of light verb structures. The study explains and evaluates the steps employed to optimize parser output in detecting open lists of light verb constructions. It discusses the qualitative usage differences of these structures in the data from the varieties and finds that ICE-GB favours fewer high frequency light verbs while ICE-IRE contains more diverse lower frequency light verbs and more passives. Overall, counts of light verb constructions are considerably higher than assumed so far. The projected counts suggest that attestations of light verb constructions will increase considerably if the search is not restricted to certain high-frequent light verbs as is typically done in studies employing manual or semi-automatic approaches to data collection.

Keywords: light verb constructions, corpus linguistics, computational linguistics, automatic parsing, collocation measures, British English, Irish English.

1. Introduction

A considerable amount of theoretical work has been carried out on the use of light verb constructions in different languages, including English (e.g. Wierzbicka 1992). Lately, a number of corpus-based studies have been conducted on the use of light verbs in Old-, Middle- and Early and Late Modern English. Recently the general interest in verb complementation patterns across varieties of English has increased (e.g. Mukherjee and Hoffmann 2009, Mukherjee and Gries 2009), as has research into methodologies in automatic extraction of verb-argument structures from corpora (e.g. O'Donnell and Ellis 2010). Nevertheless, there are only few comprehensive corpus-based quantitative studies

on light verb constructions in contemporary varieties of English, and the question of overall frequency of the construction has only received partial explanation. One such previous study is Algeo (1995), which investigates the five most frequent light verbs in LOB and Brown, another is Leech, Hundt, Mair and Smith (2009), which compares the use of *give, have* and *take* in the Brown and LOB families of corpora.

The current study aims to redress the shortage of comprehensive, quantitative corpus based research on light verb constructions in contemporary English from the British Isles. It investigates which light verb constructions are frequent in contemporary British English, and with which frequencies they are found. We consider *light verb constructions* to be those structures which are collocations of an inflectable verb, typically of low semantic specificity, with a predicate noun that in many cases is an action noun. These collocations are usually paraphrasable by a simple verb. Examples of the structures are *to make a proposal* versus *to propose*, or *to give an example* versus *to exemplify*.

The research is based on the ICE Great Britain component, and the results of searches into the light verb constructions in it are, on the one hand, compared to the situation in the larger BNC and, on the other hand, to data from ICE Ireland. We use automatically parsed versions of the corpora and extract light verb occurrences by using collocation measures (Lehmann and Schneider 2009) combined with manual filtering and then evaluate our results.

This approach enables us to identify light verb constructions in a 1 million word corpus and allows inter-variety comparison of the structures in two different ICE corpora, here ICE Great Britain and ICE Ireland. This research is valuable for determining the use

and the frequency of these collocations in contemporary English as compared to earlier varieties. It also contributes to the study of verb-object collocations in varieties of English and shows in how far different varieties of contemporary English on the British Isles differ in their usage of these constructions. The paper is structured as follows: we commence by giving a brief overview of research on light verb constructions both in English linguistics and in computational linguistics. We then continue by explaining the methodology of our approach in section 3. In section 4 we give the findings of our investigation before we offer the conclusions that we think can be drawn from the examination of our data in section 5.

2. Manual and machine-based research on light verb constructions

Light verb constructions, which are also known by various other names such as *expanded predicate* (Algeo 1995), *verbo-nominal construction* (Claridge 2000), *stretched verb constructions* or *support verb constructions* (Ronan 2012), are combinations of a semantically mostly general, and thus highly frequent, verb and a predicate noun which typically expresses the verbal process. Some researchers require these predicate nouns to be zero-derived from a verb (e.g. Wierzbicka 1982), e.g. *to have a drink*, other researchers also admit nouns which are related to a verb by other derivations (e.g. Live 1973, Algeo 1995), e.g. *to have an objection, take a decision* or *to make an apology*, while the broadest view is taken by scholars who also consider collocations where the predicate noun is not related to a verb, but the whole collocation can be paraphrased by a simple verb (Live 1973, Claridge 2000), e.g. *to take an oath* or *to cast an eye on*. In all these cases it is important that the verb-predicate noun collocation is semantically non-

compositional, i.e. that it forms a single semantic unit denoting one action. In most cases the light verb construction will have an indefinite article, but instances without an article can also be found, as can be collocations with a definite article, e.g. *to give battle* or *to have the lead* (cf. Live 1973: 36-7, Claridge 2000: 72).

The most frequently found verbs are the semantically most general ones, *have, give, take, make* or *do*, which due to their general applicability are the best-attested verbs in contemporary English in general. But also semantically more specific verbs can be found. Allerton (2002: 174-91), based on his survey of light verbs and predicate nouns starting with the letter *a,* provides an overview of light verbs with high frequency such as *have, make* or *give*, but also with medium frequency, e.g. *cause, feel, offer* or *receive*, and rare and very rare verbs, e.g. *undergo, put, arouse,* or *attract, capture* and *practise*. Though the verbs are used as the semantically 'light' element in the light verb collocation, they still contribute parts of their own meaning to the resulting collocation. Allerton (2002: 192-207) in particular points to their use in modifying causativity and aspectuality of the verb phrase, but also identifies rarer semantic contributions, such as the use of *offer* to imply tentativeness, e.g. *to offer an apology* or *to offer an answer* (Allerton, loc. cit.: 208), or verbs to denote positive and negative polarity, such as *to give* versus *to refuse admittance* (loc. cit.: 209). Still, while early English varieties fare somewhat better (Denison 1991, Brinton and Akimoto 1999, Claridge 2000, Matsumoto 2008, Ronan 2012, 2014) there remain only few comprehensive corpus-based quantitative studies on light verb constructions in contemporary varieties of English. Allerton (2002), in his book-length study, investigates light verb patterns with a complete set of light verbs in the 100 million word British National Corpus, but, as previously mentioned, the

investigation is restricted to predicate nouns starting with the letter *a* and does not consider any of the predicate nouns starting with any other letter than *a*. Therefore the study cannot offer a complete overview of the phenomenon and some of the quantitative findings may have to be adjusted when a full study becomes available. Algeo (1995), by contrast, investigates the five most frequent light verbs in the 1 million word LOB Corpus of 1960s British English and its contemporary Brown Corpus, compiled on the basis of American English. Algeo (1995: 214) finds a total of 199 constructions in Brown, with *make* having most instances (59 tokens in 44 types), followed by *have* (55 tokens in 35 types). *Give* (40 tokens, 30 types) and *take* (41 tokens, 20 types) have a middle position, *do* (4 tokens in 4 types) is hardly in evidence. In the British English LOB corpus there are a total of 245 examples, *have* is most frequent with 100 tokens in 61 types, followed by *make* (67 tokens in 37 types), *give* (40 tokens, 29 types) and *take* (38 tokens in 20 types). *Do* was not found as a light verb in the corpus. Two very interesting results emerge from these observations: first, the overall number of light verb constructions with these high-frequency verbs is higher in the British English than in the American English corpus. Second, there is a significantly higher frequency of light verb constructions with *have* in the British English corpus. The frequencies of the other light verbs remain comparable, even that of *do*, which is missing completely in LOB, but at 4 tokens is also rare in Brown.

Particularly during the last decade, general interest in verb complementation patterns across varieties of English has increased. In a comparison of first language varieties of English and second and further language varieties of English, Mukherjee and Hoffmann (2009) have shown that verb complementation patterns do vary between these

different varieties of English, and that collocation patterns do indeed grammaticalize differently in varieties of English (Mukherjee and Gries 2009, Schneider and Zipp 2013). This clearly also holds for light verb constructions. Differences have been observed in select collocations in British versus South Asian Englishes (Hoffmann, Hundt and Mukherjee 2011: 271-2) and for the use of *have* versus *take* e.g. in British versus American English (Algeo 1995). The distribution of *have* and *take*, as well as *give,* in these two varieties is further discussed by Leech *et al.* (2009). On the basis of the Brown and LOB corpora families the researchers compare the distribution of the collocations with deverbal predicate nouns derived by conversion. Due to this restriction, collocations including predicate nouns derived by other means, such as suffixation e.g. in *consideration*, are excluded. In their use with deverbal predicate nouns, frequencies of the collocations are higher in fiction than in non-fiction or press categories, and particularly high in narrative fiction as compared to fictional dialogue (Leech *et al.* 2009: 174-5). The study also confirms that *have* collocations were indeed decreasing significantly in American English to the benefit of *take* collocations, while *have* collocations experience a decrease from LOB to F-LOB, but remain higher in British English than in American English (loc. cit.: 176). A further recent study into varieties of English has been conducted by Hoffmann, Hundt and Mukherjee (2011), who investigate light verb constructions with the three high-frequency verbs *give, have* and *take* on the basis of web-derived newspaper corpora of South Asian Englishes in comparison with British English in order to determine regional specificities in the use of light verb constructions. Certain regional and inter-variety differences are determined in the study, but overall frequencies are not given.

An increasing number of corpus based studies into light verb constructions is now also carried out by using computational linguistics methodologies. Recent research is focusing on the automatic extraction of verb-argument structures from corpora (O'Donnell and Ellis 2010, Tu and Roth 2011). Tu and Roth (2011) are using a machine learning approach in order to auto-detect and extract light verb structures from corpora. They are working with structures whose predicate nouns are either nouns derived from verbs by zero-derivation, or which make use of other morphological derivations according to which the predicate noun is related to a lexical verb on the basis of NomLex (Meyers et al. 1998). They restrict their research to the light verbs *do, get, give, have, make* and *take* (Tu and Roth 2011: 34-5) and the data set is the British National Corpus. The main aim of their study is to disambiguate light verb constructions from non-light verb constructions (token-wise disambiguation). In order to allow for machine learning, a Gold Standard is manually created into which only examples with inter-annotator agreement are admitted (ibid.: 35). Tu and Roth find that candidate structures with similar surface structures (*have a look*) cannot be distinguished by statistical approaches since the latter, even though they have good results for positive recognition, have difficulties in identifying identical non-light verb constructions. Their solution to this problem is to train the classifier with contextual data that will allow it to distinguish light verb constructions from non-light verb structures. In an evaluation of Tu and Roth's approach, Nagy *et al.* (2013: 331), however, point out that the heavy restrictions set by Tu and Roth lead to only partial recovery of light verb constructions. Using the system proposed by Tu and Roth, Nagy *et al.* (2013: 333) recovered between 42% and 48% of all possible light verb constructions from English and English-Hungarian parallel corpora.

They identify this low recall rate as being due to the restriction to only some light verbs, and to a certain set of syntactic constructions, namely to verb-participle relations, verb-relative clause relations, noun-participle modification and passive constructions. Nagy *et al.* (loc. cit.) in consequence trained a system applying combined methods using statistical features, lexical features, and also morphological, syntactic, semantic and orthographical features. Their data consists of 50 Wikipedia pages annotated for different types of multi-word words, and of a likewise annotated English-Hungarian parallel corpus containing texts from different genres. They obtain precision scores of between about 59% and 63% at combined precision and recall rates (F-score) of 55% to 60% (loc. cit.: 334). This result can probably be considered the currently best available one for the automatic extraction of light verb constructions.

3. Data and Method

The data used for the current study stems from the 1 million word corpora ICE Great Britain, ICE-GB, (Nelson et al. 2002) and the similarly sampled ICE Ireland, ICE-IRE (Kallen and Kirk 2009), as well as from the 100 million word British National Corpus, BNC (Aston and Burnard 1998). Candidate constructions were extracted automatically on the basis of data that was automatically parsed with ProGres3 (Schneider 2008). The parser relies on the part-of-speech tagger and chunker LT-TTT2 (Grover 2008) and also uses the morphosyntactic information on tense, voice and aspect which LT-TTT2 provides. This method of automatic parsing was chosen as Seretan (2011) shows that collocation extraction performs better when using automatically parsed data than when observation windows are used.

3.1 A Gold Standard for the light verb *give*

In order to automatically determine the approximate number of light verb constructions in large corpora of contemporary English, and in order to be able to assess recall of automatic methods, the first step that was taken was to ascertain the exact number of light verb constructions in a sample corpus. To this end we first determined an initial Gold Standard by manually evaluating all tokens of one light verb, for which we used attestations of *give* in ICE-GB. *Give* was chosen because it is a high-frequency light verb, but easier to examine manually than light verbs with extremely high frequency, such as *have*. In total, there were 910 attestations of *give* with an object dependent in 198 different types, and these included various non-light verb collocations, as well as some questionable examples. For example, it proved to be problematic that, as also observed by Tu and Roth (2011), in some cases only manual investigation was able to distinguish the use of a predicate noun as a concrete noun in a semantically compositional context (1) and possibly (2) from true light verb uses with the same predicate noun (3), e.g.:

1. … give about three to five cytotoxic *drugs* every three weeks for about six courses (ICE GB s2a-035:2:65:A)
2. ?I can give you an oral *dose* of something (ICE GB s1a-089:2:155:A)
3. As they in turn decay they give a radiation *dose* to the lung tissues (ICE IRL W2B-040$A)

While example 1 clearly is an example of concrete use of the predicate noun, example 2 is questionable, but might be read as a light verb construction. In example 3 *give… a dose* is paraphrasable by *to dose*, as indicated by a similar example given as 4:

4. When my stomach threatened to seize up I *dosed* myself with laxatives (ICE IRL W2B-024$A)

These examples show that there is a gradient, with a grey area, of light verb and non-light verb constructions.

3.2 Automatic extraction of light verb constructions with *give*

In a second step, data of *give* collocating with an object dependent were automatically extracted from an automatically parsed version of the ICE-GB component. Light verbs and object occurrences were extracted with the following collocation methods: T-Score, O/E, χ^2, and delta P. For a detailed discussion of T-Score, O/E, and χ^2 see e.g. Evert (2009), for delta P see Gries (2013).

3.3 Evaluation of *give* in ICE GB

As the third step, we then evaluated the performance of these collocation measures on *give* against the *give* Gold Standard, which we have created in step 1. Since we created the Gold Standard for the light verb *give* in ICE-GB, we can only deliver fully automatically obtained precision and recall values for *give* light verb constructions in ICE-GB. The results plotting precision versus recall (vertical axis) can be seen in relation

to the length of the output list ranked according to the T-score collocation measure in figure 1. We use a logarithmic scale for the length of the list (horizontal axis). For example, at 20 lines of output, as illustrated in the horizontal axis, precision is 95% and recall is 10%, which means that 19 out of the 20 lines are correct, and those 19 roughly correspond to 10% of the 198 types.

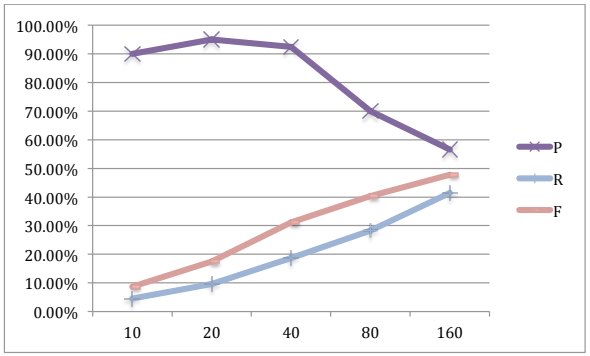

Figure 1. Precision, recall and f-measure of output according to T-Score ranking of *give* tokens in ICE-GB.

3.4 Evaluation of *give* in ICE-IRE and the BNC

As the fourth step, the same procedure was then applied to the equally sampled and similarly sized ICE Ireland in order to determine possible inter-variety variation, as well as to the British National corpus to confirm results on a larger corpus. Data from the 1 million word ICE corpora are sparse. Collocation measures are sensitive to sparse data, and thus the 100 million word BNC can be expected to suffer less from sparse data. The

result, again using T-score, is given in figure 2. As we do not have a gold standard for the BNC, we must point out that recall cannot reach 100% as some light verb constructions seen in ICE-GB remain unseen in the much larger BNC, and precision values on long output lists are considerably too low as increasingly more light verb constructions that would be annotated as correct on manual inspection are unseen in the much smaller ICE-GB.

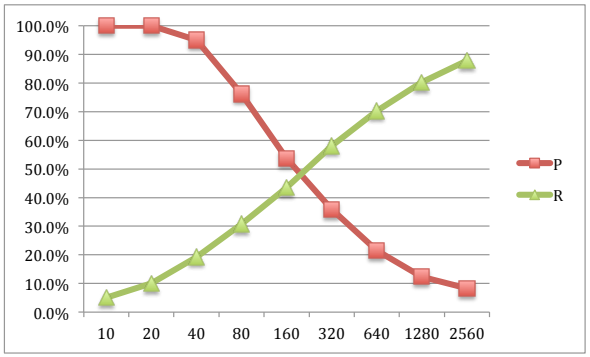

Figure 2. Evaluation of *give* precision and recall on the BNC using T-Score and a simple filter

3.5 Using WordNet and NomLex

As these results still contain too many false positives, which leads to too low precision when using longer lists, we applied semantic specification methods in the fifth step, in order to restrict the number of false positives. The first device used is WordNet (Fellbaum 1998), which is a detailed hierarchical semantic lexicon and contains among

other taxonomies 44 simple classes (the so-called lexicographer files), 24 of them for nouns. From these we selected those noun classes in which nouns predicted to be appropriate for light verb constructions were found. They are the following classes:

> nouns denoting acts or actions (class 04), attributes of people and objects (cl. 07), body parts (cl. 08), cognitive processes and contents (cl. 09), communicative processes and contents (cl. 10), natural events (cl. 11), feelings and emotions (cl. 12), goals (cl. 16), possession and transfer of possession (cl. 21), natural processes (cl. 22), relations between people and things or ideas (cl. 24), stable states of affaires (cl. 26).

We decided to exclude the remaining noun classes denoting

> Animals (class 05), man-made objects (cl. 06), food and drinks (cl. 13), groupings of people or objects (cl. 14), spatial position (cl. 15), natural objects (cl. 17), people (cl. 18), natural phenomena (cl. 19), plants (cl. 20), qualities and units of measure (cl. 23), two and three dimensional shapes (cl. 25), substances (cl. 27), and time and temporal relations (cl. 28).

Further, we used NomLex (Meyers et al. 1998) in order to automatically detect nouns which are related to verbs. NomLex stores verb-noun alternations as illustrated by the extract given in table 1.

```
nomlex( abandonment, [abandon, for, to ]).
nomlex( abasement, [abase ]).
nomlex( abatement, [abate ]).
nomlex( abbreviation, [abbreviate, to ]).
nomlex( abdication, [abdicate, to ]).
nomlex( abduction, [abduct, from ]).
nomlex( abductor, [abduct, from ]).
nomlex( aberration, [err, on, in ]).
nomlex( abhorrence, [abhor ]).
nomlex( ability, []).
nomlex( ability, []).
nomlex( abjuration, [abjure, for, to ]).
nomlex( abolishment, [abolish ]).
nomlex( abolition, [abolish ]).
nomlex( abomination, [abominate ]).
nomlex( abortion, [abort, from ]).
nomlex( abrasion, [abrade ]).
nomlex( abrasive, [abrade ]).
```



```
nomlex( abrasive, []).
nomlex( abridgement, [abridge ]).
…
```

Table 1: examples verb-nominalization alternations stored in NomLex.

We first used the above resources as a filter: we discarded verb-predicate noun combinations appearing in the ranked lists with a predicate noun belonging to a non-licensed WordNet class, and/or nouns not appearing in NomLex. Either of these approaches led to higher precision, but drastically lower recall. This indicates that lexical resources like NomLex are far from being complete. It also entails that approaches which treat the data obtained by the application of such resources as Gold Standards compare to an in fact unrealistic Gold Standard. These are the difficulties also seen in Tu and Roth (2011), who use WordNet and NomLex as filters. They "filter out approximately 55% potential negative examples" (Tu and Roth 2011: 35), which is a permissible operationalization when interested in token-wise disambiguation and not in full recall of all light verbs. As a filtering approach remains too restrictive for our needs, we switched to a weighting approach, which slightly punishes, i.e. reduces, collocation scores instead of completely discarding candidates that are not licensed. This method led to an increase in F-score of about 1%.

3.6 Extending to all active voice light verbs

In the sixth step, we have extended from the light verb *give* to all verbs in the active form. In order to measure precision of our approach in the sixth step, we manually annotated the most highly ranked output of all active verbs to find all the accurate examples of light verb constructions contained in the output list.

We have applied T-Score sorting to all the object relations that the tagger reported in the corpus data. In the following two tables we give the first 20 lines only of the total output lists. It is the top lines which usually report the most confident results, accuracy decreases in the lower lines of the parser output. Data from on ICE-GB is shown in table 2, and data from the BNC is shown in table 3. These results still contain incorrect parses and in order to weed these out, we have used the weighting approach from the previous step, filtered hapax legomena, occurrences with the verb *be*, object pronouns , and *errm* as object. Allocating object status to *errm* is a frequent parsing mistake from the spoken section of the corpora. Further frequent false positives contain indefinite object pronouns like *something* or *anything*, which are found both in ICE-GB and the BNC, or direct objects like *thing* or *pp*, i.e. 'pages'. As indicated in table 3, we have annotated the candidates that we consider correct with a '+' in the last column. Such annotations allow us to measure the precision of the parser output: 12 correct lines out of 20 correspond to 60% correctness.

4447	OE	T	Chi	V	Object	f	f(V)	f(N)
icegb	27.37	9.919	2775.21	take	place	106	1231	218
icegb	6.67	6.011	521.354	have	effect	50	4219	123
icegb	3.20	5.881	268.545	do	something	73	2775	568
icegb	3.00	5.814	264.213	do	thing	76	2775	632
icegb	16.69	5.481	545.332	make	decision	34	1440	98
icegb	7.44	5.406	493.234	have	look	39	4219	86
icegb	38.86	5.246	1111.84	take	care	29	1231	42
icegb	6.69	5.174	203.018	say	thing	37	606	632
icegb	129.9	5.156	3470.19	pay	contribution	27	240	60
icegb	11.18	5.069	386.333	get	have	31	2432	79
icegb	5.18	5.039	259.92	do	work	39	2775	188
icegb	27.89	5.009	708.861	play	part	27	243	276
icegb	54.39	4.908	1319.05	come	home	25	245	130
icegb	17.63	4.809	509.008	get	credit	26	2432	42
icegb	48.88	4.798	1133.77	play	ball	24	243	140
icegb	3.57	4.667	356.966	have	idea	42	4219	193
icegb	25.10	4.604	540.092	ask	question	23	345	184
icegb	2.86	4.599	348.832	have	problem	50	4219	287
icegb	3.54	4.595	209.002	do	anything	41	2775	289
icegb	155.06	4.553	3226.11	answer	question	21	51	184

Table 2: Results of T-score sorting of the parser output on ICE-GB.

117249	OE	T	Chi	V	Obj	f	f(V)	f(N)	manu
bncx	22.40	97.07	219641	take	place	10325	128201	21501	+
bncx	8.47	64.00	52584	have	effect	5266	303211	12254	+
bncx	272.5	59.53	968964	shake	head	3570	6221	12594	
bncx	52.77	55.38	167118	see	pp	3187	112461	3212	
bncx	5.24	55.03	22132	do	thing	4626	157530	33518	
bncx	41.21	53.51	119208	ask	question	3008	30365	14376	+
bncx	92.01	49.44	226112	play	role	2499	19223	8451	+
bncx	32.78	47.98	76299	play	part	2450	19223	23253	+
bncx	6.31	47.22	18094	take	part	3148	128201	23253	+
bncx	7.09	47.08	21023.8	do	anything	3004	157530	16078	
bncx	31.47	46.45	68682.6	go	home	2302	26840	16301	
bncx	4.65	46.34	15831.8	do	something	3484	157530	28412	
bncx	12.76	46.23	31919.8	make	sense	2516	147869	7971	+
bncx	7.85	46.16	21934.2	do	job	2798	157530	13522	+
bncx	12.73	45.65	31139.6	make	decision	2455	147869	7801	+
bncx	142.24	45.31	293114	open	door	2083	11317	7740	
bncx	4.54	44.51	25899.6	have	idea	3257	303211	14139	+
bncx	159.3	43.15	297871.0	answer	question	1886	4923	14376	
bncx	6.37	43.03	28310.3	have	look	2605	303211	8059	+
bncx	10.93	42.48	24204.3	make	use	2187	147869	8088	+

Table 3: Results of T-Score sorting on BNC, with results deemed correct marked by a +.

We have manually annotated the first 320 lines of output of the t-score ranked list in the manner indicated in table 3. Precision was about 50%, and there additionally were many correct light verb constructions further down in the list. In order to assess precision lower down in the list, we have used a stratified approach: at list position 2000, 5000 and 50,000, we manually annotated 100 lines and calculated their precision. The development of precision is shown in figure 4, in the red line. As many of the verbs appearing in the lists (for example *see* in table 3) very rarely participate in light verb constructions, we have created a list of those verbs which appeared as light verbs with several object nouns. The list is given in figure 3.

take, have, do, ask, play, make, give, provide, draw, tell, pay, meet, change, keep, form, attend, raise, become, reach, cause, live, sing, turn, catch, perform, adopt, put, cover, lead, send, focus, show, receive, suffer, issue, exercise, pay, form, set, feel, ring, issue, suffer, commit

Figure 3. List of 44 permitted verbs for the verb-restriction experiment.

Precision results for using the restriction to these 44 verbs are given in figure 4 in the blue line. The output line number (vertical axis) was normalized to the full verb output, indicated by the red line, in order to be able to compare the results. As can be seen in the figure, the restriction to known light verbs improves precision considerably at all observed levels.

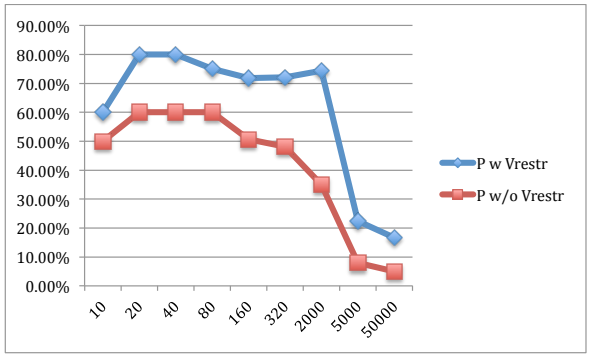

Figure 4: projected extension of the top 320 verbs from the BNC to overall data.

3.7 Light verb constructions in the passive

To determine the overall number of light verb constructions, we have used active instances of these constructions with the verb *give* to build our gold standard, and the evaluation and reported results are also done using active constructions.

However, in addition to active voice, we of course also find ample evidence of passive verbs in the corpora, and their contribution to the overall number of light verbs should also be taken into account. In order to determine the numbers of potential light verb constructions in the passive in our corpora, we need to determine their structural specificities. Structurally, light verb constructions in the passive involve subject and verbs while light verb constructions in the active involve verbs and an object.

We did not include passive verbs into all our investigations, instead we have carried out steps 2 to 6 (sections 3.2 to 3.6) separately also for passive verbs. Figure 5 shows the top 20 candidates according to the T-score collocation measure from the BNC. Precision for these is 19/20 = 95%, until position 100 it drops to 56/100 = 56%. As we have no gold standard for passive constructions, we cannot assess recall.

	O/E	T	Chi	pass-V	Subj	f	f(V)	f(N)	manu
bncx	23.17	29.03	95818.3	make	attempt	858	31144	1460	+
bncx	11.05	28.66	44670.9	make	decision	853	31144	3043	+
bncx	19.37	27.41	71219.6	take	decision	768	15999	3043	+
bncx	14.36	24.62	42574.1	do	work	624	12605	4233	+
bncx	48.00	23.13	125425.8	hold	meeting	540	10189	1356	+
bncx	23.11	22.55	57511.3	take	action	518	15999	1720	+
bncx	14.96	20.82	31968.2	make	effort	446	31144	1175	+
bncx	163.7	20.32	329449.5	reach	agreement	414	2038	1524	+
bncx	23.66	20.31	47980.4	make	progress	420	31144	700	+
bncx	35.54	20.20	70911.9	take	care	413	15999	892	+
bncx	53.84	19.87	103783.5	hold	election	398	10189	891	+
bncx	39.63	19.79	75904.7	take	step	396	15999	767	+
bncx	10.68	19.51	20097.3	make	application	396	31144	1462	+
bncx	42.93	19.09	76310.9	ask	question	368	5269	1998	+
bncx	16.69	18.99	29672.2	make	payment	370	31144	874	+
bncx	21.10	18.91	37152.2	make	reference	365	31144	682	+
bncx	60.82	18.64	103111.9	pay	attention	350	5622	1257	+
bncx	9.01	18.38	15103.6	make	order	354	31144	1548	+
bncx	27.63	17.53	41439.74	build	house	312	6619	2095	
bncx	12.36	17.17	18021.49	make	provision	305	31144	973	+

Figure 5. Marked-up top candidates for light verb constructions from BNC

In ICE-GB and ICE-IRE, the frequency at the 100th position is 2. We have stopped evaluating at this position. This gave us an additional 138 light verb construction tokens in ICE-GB and 189 in ICE-IRE. The vast majority of the light verb construction types in the passive were already in our Gold Standard (i.e. we have recognized them with our approach on active verbs), but we discovered some additional light verb types, types which mainly seem to be used in the passive. We added the top-scoring 100 passive constructions according to T-score from BNC, ICE-GB and ICE-IRE to our Gold Standard, for the items which we manually accepted as correct, and which appear at least once in at least two corpus. The list is given in figure 6.

Corpus	V	Subj	f
BNC	make	application	396
BNC	carry	work	279
BNC	make	comparison	235
BNC	do	damage	182
BNC	make	mention	174
BNC	give	notice	163
BNC	take	measure	144
BNC	carry	study	140
BNC	make	announcement	138
BNC	make	appointment	138
BNC	make	contract	137
BNC	carry	research	130
BNC	adopt	approach	123
BNC	make	allowance	123
BNC	make	adjustment	116
BNC	provide	example	115
BNC	undertake	work	113
BNC	make	improvement	112
BNC	carry	test	108
BNC	reach	decision	102
ICE-GB	pay	contribution	3
ICE-GB	take	measurement	3
ICE-GB	exert	pressure	2
ICE-GB	conduct	test	2
ICE-GB	perform	analysis	2
ICE-IRE	make	submission	4
ICE-IRE	make	assumption	4
ICE-IRE	place	advertisement	3
ICE-IRE	perform	abortion	3
ICE-IRE	make	accusation	3
ICE-IRE	make	award	3
ICE-IRE	make	copy	3

ICE-IRE	afford	protection	2
ICE-IRE	perform	comparison	2
ICE-IRE	undertake	exercise	2
ICE-IRE	levy	charge	2

Figure 6. Manually accepted light verb constructions added to the Gold Standard

3.8 Which collocation measure?

In the next step, we compared the usefulness of several collocation measures in determining correct light verb constructions within the verb-object collocations in the output of the parser. For a comprehensive overview of collocation measures see Pecina (2009). We carried out a sort with the measures T-Score, O/E, χ^2, delta-P. We first used O/E without a T-Score filter. In the results we found that the output was dominated by rare collocations and parse errors. Therefore we used a T-Score filter to erase rare collocations. The results of this step using a T-Score filter T>15 are illustrated in fig. 7, where the collocations in the output that are considered as correct light verb constructions are marked by '+'.

578	OE	T	Chi	V	Obj	f	f(V)	f(N)	manu
bncx	3630.7	15.55	878363	beg	pardon	242	1066	374	+
bncx	1097.63	18.50	376046	commit	suicide	343	3320	563	+
bncx	1023.84	16.47	278060	light	cigarette	272	1383	1149	
bncx	868.23	15.82	217540	celebrate	anniversary	251	2405	719	
bncx	797.01	16.01	204437	clear	throat	257	2653	727	
bncx	743.37	18.46	253726	ring	bell	342	3185	864	
bncx	705.75	18.41	239441	press	button	340	3139	918	
bncx	633.74	15.56	153676	bear	resemblance	243	5649	406	+
bncx	584.87	15.00	131804	score	try	226	2337	989	
bncx	519.30	17.71	163058	pose	threat	315	1641	2211	+
bncx	500.00	16.09	129577	exert	influence	260	914	3403	+
bncx	492.65	15.93	125227	mark	anniversary	255	4306	719	
bncx	478.24	18.61	165853	sing	song	348	2143	2031	+
bncx	421.54	20.71	180984	wait	minute	431	1962	3117	
bncx	417.76	24.12	243472	score	goal	585	2337	3584	+
bncx	414.72	17.13	121865	earn	living	295	4242	1003	
bncx	405.00	18.03	131902	speak	english	327	4038	1196	
bncx	341.01	19.10	124535	plan	permission	367	3721	1730	

```
bncx    318.33      16.74    89284   commit    offence      282    3320   1596 +
bncx    316.63      16.41    85396   catch     glimpse      271    7745    661 +
```

Figure 7: Sort by O/E, with a high T-score filter, correct examples marked by +.

Even though a number of false positives are reported by this search, this measure yields interesting results in that it finds rare collocations, and some very rare support verbs are amongst these, such as *beg* or *pose*. The measure is good for recall, but low on precision (9/20 in the top ranks.). Similar results were reached in a search with Delta P, which offered nice rare collocations, with some rarer objects among them. Even though this search provided generally good recall, it proved to be similar to the measure O/E. A comparison of the tested collocation measures can be found in figure 8. Here the investigated data was the top 160 lines of parser output, T was set at > 2, the NomLex weighting factor, as well as the WordNet weighting factor, were set at 50%. By this measure, nouns that were not part of these lists were punished in the weighting, but not discarded.

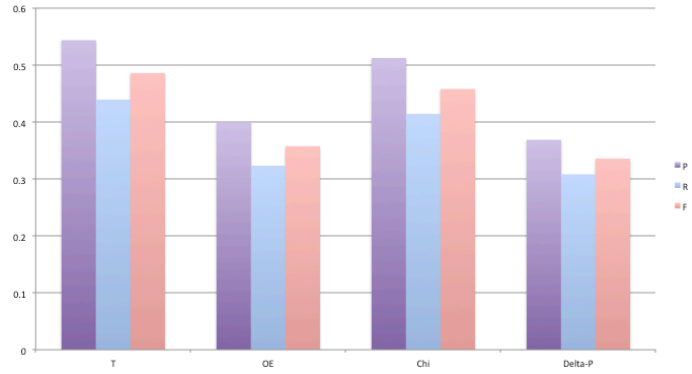

Figure 8: P/R/F of different collocation measures on the top 160 lines of parser output.

On the basis of the output evaluations of the parsed material we found that the most suitable measure amongst those tested to determine correct light verb and predicate noun constructions is T-score, which we then proceeded to use for the automatic extraction of light verb constructions.

4. Results

4.1 General results

4.1.1 Light verbs in the active voice

In order to evaluate the results of the corpus searches from British and Irish English, we created lists of the relevant light verb constructions in both varieties. These were created on the basis of the Gold Standard (see section 3). First, this consisted of the manually evaluated top 400 results from the BNC, followed by intermediate 100 results at 2000;

4000; 6000; and 80.000 verb-noun combinations. These manual evaluations yielded 255 correct types. Second, the exhaustive list of *give* attestations taken from ICE-GB was added to the Gold Standard, these yielded 198 types. On the basis of these we computed the attestations and the attestation differences in and between ICE-GB and ICE-IRE. We carried out two searches. On the one hand we searched for items by T-score. On the other hand we searched for frequencies of light verb constructions, here we restricted the search to only display those items which have a frequency f > 2**.** While frequency (and its differences between ICE-IRE and ICE-GB) is easy to interpret, T-score (and its difference) has the advantage that it is based on statistical significance, thus punishing chance events.

Figure 9 shows the tokens which have higher T-scores in Irish English than in British English. We see e.g. that the difference between the T-scores for *ask question* is 2.22 between ICE-IRE and ICE-GB. Using frequency differences, we find that *ask question* is found 26 times as frequently in ICE Ireland as in ICE Great Britain (cf. example 6). The next more frequently used items in ICE Ireland than in ICE Great Britain are *give advice* (example 7), *do work, give view* and *make decision* (examples 8, 9 and 10 respectively) which are between 23 and 20 times as frequent in ICE Ireland as in ICE Great Britain.

6. Uh I think it it would be fairer if a spokesperson from each of the parties in opposition were allowed to *ask* a final *question* (ICE-IRE S1B-056:1:63:J)
7. An animal nutritionist has to be capable of *giving advice* on a multitude of different areas (ICE-IRE W2B-024:2:4)

8. I m just *doing* some *work* here transcribing tapes (ICE-IRE S1A-091:1:5:B)
9. We will have systems in place (for example a website for *giving* your *views*) that allow you to contribute to our policies effectively (ICE-IRE W2D-005:2:54)
10. They wouldn't *make* a unanimous *decision* (ICE-IRE S1A-030:1:117:C)

Items about 5 to 8 times more frequent in ICE-IRE than in ICE-GB are the *make* collocations *make statement* and *make sense* (examples 11 and 12) as well as the *have* collocations *have regard, have interest* and *have difficulty* (examples 13, 14 and 15).

11. I *made a statement* in this House yesterday regarding the circumstances of the appointment of Mr Harry Whelehan to the presidency of the High Court. (ICE-IRE S1B-057:1:32:G)
12. It is thus by reference to the specifics of geographical locality that I hope to *make sense* of the surprisingly different responses of orthodox Calvinists in Belfast and Princeton (…) (ICE-IRE W2A-003:1:13)
13. In determining dividend policy the Directors will *have* particular *regard* to the stability of the gross dividend taking into account the related tax credit. (ICE-IRE W2D-007:1:100)
14. Uh indeed uh I can recall uh that I *had* a lot more *interest* in a a young girl who was only a few months younger than me than I had in politics. (ICE-IRE S1B-055:1:41:F)
15. And they re *having difficulty* getting constructive ball out of this half of the field. (ICE-IRE S2A-006:1:19:B)

We initially noted a particularly high incidence of *make*-based collocations in the highly frequent collocations in the Irish English data. Further collocations involving *make* which are between twice and four times as frequent in ICE-IRE are *make effort, make comment, make choice, make contribution, make profit, make use, make sound* and *make order*. However in significance testing these differences in frequency between ICE-IRE and ICE-GB did not appear as statistically significant. Here research with data from larger corpora might add further interesting insights as to whether there are after all significant differences in the use of *make*-based collocations between the two varieties.

In correspondence with the situation in ICE Ireland, we also found items to be considerably more frequent in ICE Great Britain than in ICE Ireland (Figure 9). The items with the highest frequency in ICE-GB were *have experience* and *take place*, which were 12 times more frequent in ICE-GB than in ICE-IRE. *Give information* and *take action* as well as *make change, have lunch* and *give evidence* were 11 and 10 times as frequent in ICE-GB (view examples 16-20).

16. If you do not *give* this *information* the optician can refuse to test your sight (ICE-GB W2D-001:1:24)
17. It emerged however that the word restraint used by London and Washington did not necessarily mean *taking* no *action*. (ICE-GB S2B-015:1:28:C)
18. In no sense where necessary and if necessary would I be afraid to *make changes* in government policy. (ICE-GB S2B-003:1:31:C)
19. And I *had lunch* with someone Monday so. (ICE-GB S1A-055:1:140:A)

20. This *gives evidence* of a method of perpetuating tradition and distribution of news. (ICE-GB W1A-002:1:55)

Other *have* and *take* collocations were also noted to be considerably more frequent in ICE-GB than in ICE-IRE. Between 4 and 8 times more frequent were the collocations *have influence, have trouble, have doubt, have reason, have go, have feeling, have chance, have history, have access, have impact, have look* and *have holiday,* as well as *take advantage, take care* and *take view.* We further find *take decision* to be 3 times as frequent in ICE-GB as in ICE-IRE, which may be explained by the preference for *make decision* over *take decision* in the Irish data set. Figure 9 gives an overview of light verb constructions and of how much more frequently they are used from the point of view of the Irish English data, figure 10 shows those light verb constructions which are more frequent in ICE Great Britain than in ICE Ireland, and how much more frequent they are than their Irish counterparts.

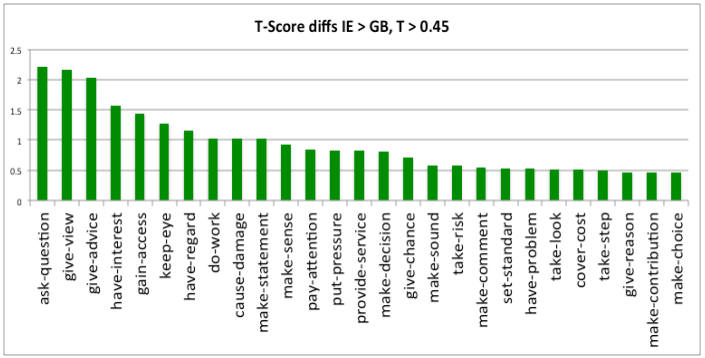

Figure 9: Light verb constructions with higher frequencies in ICE-IRE than in ICE-GB

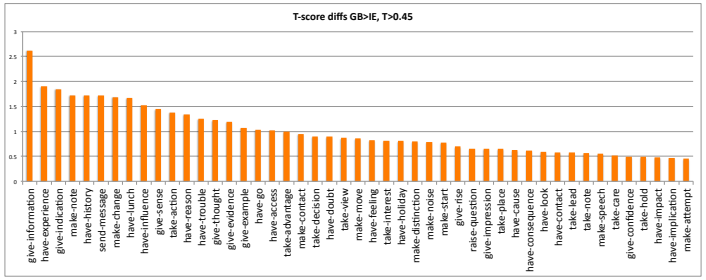

Figure 10: Light verb constructions with higher frequencies in ICE-GB than in ICE-IRE

It should be reiterated that direct comparison of the data in ICE Great Britain and ICE Ireland is made more difficult by two specificities of the research settings: on the one hand the Gold Standard has been developed on the basis of ICE Great Britain; this has the effect that examples which are only attested in ICE Ireland or in the British National Corpus may be missed. On the other hand rare verbs may be underrepresented in this approach as they are at risk of being filtered out by the triangulation process, which requires that the verbs be present in all three corpora. In spite of these restrictions, our approach was still able to obtain rare light verb collocations, which are typically not considered in other automated studies. This applies to the light verb constructions *play part*, *provide finance* or *catch glimpse* as in:

21. Both have been assured by Teheran that the jets will *play* no further *part* in the war. (ICE-GB S2B-018:1:21:A)
22. (…) he had agreed together with his father to *provide finance* for Walling's visit to Africa (ICE-GB S1B-068:2:116:B)
23. Yet if we should *catch a glimpse* of the fake plastic beams the mass-produced horsebrasses cartwheels and hunting prints there will often as not be unnerving claims of genuine olde worlde atmosphere for this mishmash of bogus antiquity. (BNC A0B:112)

Play part can be seen as a light verb construction that corresponds to *participate*, *provide finance* is paralleled by *to finance*, while *catch glimpse* corresponds to *glimpse*. In all

three cases the replacement of the light verb construction by the simplex would have been possible without extensive meaning changes, which indicates light verb status.

4.1.2 Light verbs in the passive voice

According to Leech *et al.* (2009:150), the overall frequencies of passives in the 1960ies LOB corpus were 11.8%, and have risen to 13.4% the 1990ies F-LOB. For our *International Corpus of English* data sets, we have determined that in ICE-GB, 13.8% of the verbal groups are in the passive, the corresponding number in ICE-IRE is 12.7% Out of these, following the approach described in section 3, we extracted 138 passive light verb constructions in ICE-GB (8.5% of all light verb constructions) and 189 in ICE-IRE (11.7% of all light verb constructions). These percentages of light verb constructions in the passive are significantly higher in ICE-IRE at a significance level of $p < 0.0027$ according to chi-square contingency test with Yates' continuity correction.

As with the attestations in the active voice, we were further able to observe differences of the attestation patterns in the British and in the Irish corpora. Different usage patterns, based on T-score differences, are given in figure 11. We have set T-score >= 0.2, and f > 1.

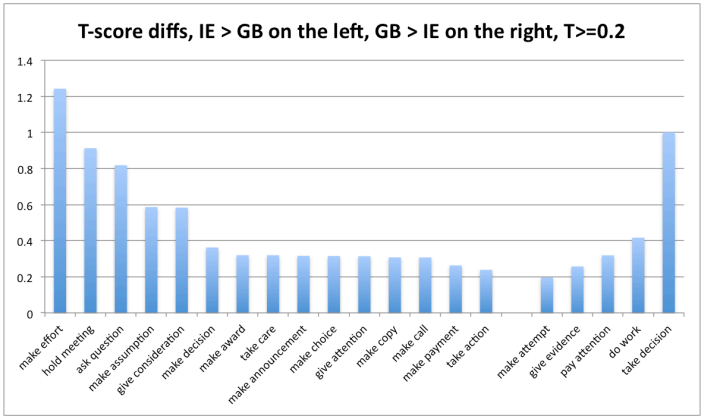

Figure 11. T-score differences of passive LVCs. ICE-IRE overuse on the left, ICE-GB overuse on the 5 rightmost items.

In the attestations of light verbs in the passive, we can observe tendencies that are similar to those in active use. In the Irish passive data, we again notice the high instance of *make* collocations, as in *make effort* (versus *take effort*), *make assumption* or *make decision*. But also *questions* are *asked* more frequently, more *care* is *taken,* and more *consideration* is *given* in the ICE-IRE data, e.g.

24. Which leads me to suggest that particular *care* should *be taken* to avoid gratuitous insults … (ICE-IRE W1B-028:2:33)

By contrast, ICE-GB in the passive data, like in the active data, comparatively overuses *make attempt* and *give evidence. Take decision* is even considerably more clearly overused in the passive in ICE-GB, as in

25. As far as Watson was concerned you had his interviews when the *decision* to seize the vessel *was taken* on the twenty-second of August (ICE-GB S1B-063:1:63:B)

In the passive data the differences between *take* and *make decision* are statistically significant in ICE-GB and ICE-IRE. In contrast to ICE-IRE, *do work* dominates in passive light verb constructions in ICE-GB, while in ICE-IRE strongly favours *do work* in active light verb constructions, where the collocation is about 20 times more frequent than in ICE-GB in the active voice. Compare example 26 to example 8 above:

26. But they recommend that *any work* by the water electricity and gas authorities should *be done* before the scheme is started. (ICE-GB W2C-017:4:87)

4.2 Specific results

When comparing the overall frequencies of the active and passive collocations that have been identified as light verb constructions in ICE Great Britain and ICE Ireland, we find that the overall difference in the number of light verb constructions is statistically significant at $p < 1\%$ according to the chi-square contingency test. This also holds when removing all verbs whose expected (E) values are below 5, as is recommended for chi-square testing. That ICE-GB contains slightly more light verbs may partly be explicable

by the fact that the development of the Gold Standard on the basis of ICE Great Britain is likely to have caused under-attestation in ICE Ireland. Overall and specific light verb counts in those two corpora are shown in table 4.

	GB active	IRE active	GB passive	IRE passive	GB	IRE	Total
adopt	2	6	0	3	2	9	11
afford			0	2	0	2	2
ask	23	49	2	5	25	54	79
attend	5	7			5	7	12
carry	10	5	3	10			
catch	5	2			5	2	7
cause	8	14	0	4	8	18	26
commit	2	2			2	2	4
conduct			2	2	2	2	4
cover	3	5			3	5	8
do	45	56	9	7	54	63	117
draw	8	10	3	5	11	15	26
Exert			2	0			
express	2	6	2	0	4	6	10
feel	3	2	0	2	3	4	7
form	7	6	0	2	7	8	15
gain	3	10			3	10	13
give	228	193	16	21	244	214	458
have	483	407			483	407	890
hold	2	4	3	9	5	13	18
issue	2	5			2	5	7
keep	7	15			7	15	22
live	8	10			8	10	18
make	271	260	58	80	329	340	669
meet	9	12	0	3	9	15	24
pay	12	17	6	2	18	19	37
perform			2	2	2	2	4
place	2	3	0	3	2	6	8
provide	32	40	2	3	32	43	75
put	7	12	2	0	9	12	21
raise	7	4	0	5	7	9	16
reach	2	4	2	2	4	6	10
receive	11	10			11	10	21
run	2	3			2	3	5
score	2	3	2	0	4	3	7

send	12	3			12	3	15
set	3	5			3	5	8
show	6	7	0	3	6	10	16
take	245	217	22	12	267	229	496
undertake	3	2	0	2	3	4	7
Total	1472	1411	138	189	1603	1590	3193

Table 4: Counts of light verb constructions in ICE-GB and ICE-IRE.

As indicated above, the distinction between *make* and *take* is not statistically significant, neither is the observable distinction between *make decision* and *take decision* significant in the active voice. In view of Algeo's (1995: 203-17) and Leech *et al.*'s (2009: 179) results, the differences in the use of *have* versus *take* might have provided an interesting insight into whether the Irish data patterns rather with British or rather with American English. However, overall no specific patterning of ICE Ireland with regard to the other two varieties can be confirmed.

We can, however, make statements on frequencies of other light verbs. For an overview of observed frequencies, view table 5, showing differences in attestations of specific light verbs in ICE Great Britain and ICE Ireland. The table contains the relative overuse of ICE-IRE as compared to ICE-GB in the counts above the bar and the relative underuse of ICE-IRE as compared to ICE-GB below the bar. In order to reduce the effect of chance events, we have removed rare verbs, i.e. all verbs occurring less than 10 times in total (which corresponds to E values below 5, as typically done in chi-square testing).

V	GB	IRE	∑	IE/GB
adopt	2	9	11	4.5
gain	3	10	13	3.3333
hold	5	13	18	2.6

cause	8	18	26	2.25
ask	25	54	79	2.16
keep	7	15	22	2.1428
meet	9	15	24	1.6666
show	6	10	16	1.6666
express	4	6	10	1.5
reach	4	6	10	1.5
attend	5	7	12	1.4
draw	11	15	26	1.3636
put	9	12	21	1.3333
raise	7	9	16	1.2857
provide	34	43	77	1.2647
live	8	10	18	1.25
do	54	63	117	1.1666
carry	13	15	28	1.1538
form	7	8	15	1.1428
pay	18	19	37	1.0555
make	329	340	669	1.0334
receive	11	10	21	0.9090
give	244	214	458	0.8770
take	267	229	496	0.8576
have	483	407	890	0.8426
send	12	3	15	0.25
∑	1585	1560	3143	0.9842

Table 5: frequencies of light verbs in ICE-GB and ICE-IRE ordered by frequency ratio

We can observe that those verbs which are used with high frequency, *have, take* and *give* are more frequent in ICE-GB than in ICE-IRE. To a lesser extent this also holds for *make*. By contrast, light verbs with lower frequency tend to be more frequent in ICE-IRE. This is true for a large range of verbs, including *ask, keep* and *gain*. The list of 20 verbs of ICE-IRE overuse is matched by a list of only 5 verbs of ICE-GB overuse. This finding indicates that in the British English data, there is a higher concentration in the use of fewer light verb types. By contrast in the Irish data we have less focus on the small set of high-frequency light verbs. This insight also explains why the list of new passive light

verb constructions found in ICE-IRE was much longer than in ICE-GB (figure 6 in section 3.7).

Finally, we would like to return to the question of the overall token frequencies of light verb constructions in the ICE corpora under investigation. As shown by the sums given in table 4 above, the estimated token frequencies in the active voice are indicated at an average of about 1450 active plus 160 passive light verb constructions per one million words resulting in an approximate average of about 1600 light verb constructions per 1 million word corpus. These counts are restricted by the triangulation process which has been applied to the active voice, according to which rare words have to be present in both ICE-GB and ICE-IRE, as well as in the BNC. This bias in favour of British data potentially over-represents English collocations over specific Irish collocations, which might have been filtered out by the process because they were not found in the British English data sets and were therefore lost to triangulation. Bearing this in mind, the expected total of light verb constructions in the corpus should be above 1600 tokens.

In the BNC, the minimum counts of light verb constructions consists of all manually accepted 255 types, which are evidenced in 189,917 tokens, i.e. 1900 tokens per million words, plus a minimum of 20 passive types in 3168 tokens. Here it needs to be borne in mind that the types are Gold Standard based. In comparison with Algeo's (1995) counts of 199 tokens with select light verbs for Brown and 245 tokens in LOB the difference is striking. These projections on the basis of the automatic extraction of data from the BNC and ICE-IRE and ICE-GB show that the overall numbers of light verb constructions in these corpora of contemporary English are considerably higher than has

so far been assumed on the basis of the manual or semi-automatic investigation of only select high-frequency light verbs.

5. Conclusions

In this study we have employed an automatic parser in order to determine the overall amount of light verb construction in two varieties of Present Day English, British English and Irish English. We have used an inclusive approach which restricted neither the number of allowed light verbs nor the morphological pattern of the predicate nouns. With respect to determining the best methodology, we found that the T-score collocation measure seems to work better than other methods, and that the use of NomLex and WordNet was only helpful when weighting was introduced. This confirms that a Machine Learning approach, as e.g. Nagy (2013) has used, can lead to better results, since the various features are weighted automatically in Machine Learning approaches.

Our research has shown that the extracted frequencies of attested light verb constructions are around 1600 per 1 million word corpus, with slightly fewer examples in ICE-IRE than in ICE-GB, but these must be seen as minimum figures due to the methodological constraints of the automated approach. Even these minimum extracted figures, however, are already considerably higher than any counts that have previously been obtained by researchers who have attempted to determine overall counts of light verb constructions in comparable corpora by focusing on select, high-frequency light verbs only. We were able to obtain an open list of light verb constructions, and this leads to the observation that considerably larger frequencies of light verb constructions can be posited for these varieties of English than previously assumed. In order to determine the

exact frequency of light verb constructions in the corpora, even more parser optimization will, however, be required in the future.

Some overall differences in frequencies of attestations between ICE-GB and ICE-IRE were observable. Most notably, ICE-IRE uses significantly more light verb constructions in the passive than ICE-GB does. We have further observed differences in ranking of light verbs in the British and Irish English corpus data: ICE-GB shows higher frequencies of fewer highly frequent light verbs, whereas the collocations in ICE-IRE evidence higher versatility in the light verbs used. It would be interesting to determine whether there is an ongoing process of increasing restriction of light verbs to only high-frequent light verbs in British English. In order to assess this, parallel corpora representing different time periods of contemporary English, such as Brown, Frown and BBrown or LOB, FLOB and BLOB could profitably be tested by these automated methods to maximize both precision and recall. In further research, we would like to extend our approach to include prepositional phrases and alternation frequencies in order to improve the recall of our automated approach and thus obtain even more comprehensive results.

6. References

Algeo, J. 1995. 'Having a Look at the Expanded Predicate.' In: Aarts, B. & Ch. F. Meyer (eds) *The Verb in Contemporary English: Theory and Description.* Cambridge: CUP. 203-17.

Allerton, D. J. 2002. *Stretched Verb Constructions in English*. London: Routledge.

Aston, G. and L. Burnard. 1998. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.

Claridge, C. 2000. *Multi-word Verbs in Early Modern English*. Amsterdam: Rodopi.

Evert, S. 2009. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook, article 58*. Mouton de Gruyter, Berlin.

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database. MIT Press.*

Gries, S. T. 2013. '50-something years of work on collocations: what is or should be next …'. *International Journal of Corpus Linguistics* 18:1, 137-166.

Grover, Clare. 2008. *LT-TTT2 Example Pipelines Documentation.* Edinburgh Language Technology Group. Edinburgh: University of Edinburgh.

Hoffmann, S., M. Hundt and J. Mukherjee. 2011. 'Indian English – an Emerging Epicentre? A Pilot Study on Light-Verbs in Web-derived Corpora of South Asian Englishes.' *Anglia 12:3-4,* 258-280.

Leech, G., M. Hundt, Chr. Mair and N. Smith. 2009. *Change in Contemporary English*. Cambridge: Cambridge University Press.

Lehmann, H. M. and G. Schneider, 2009. 'Parser-Based Analysis of Syntax-Lexis Interaction'. In: Jucker, A. H., D. Schreier and M. Hundt (eds) *Corpora: Pragmatics and Discourse: Papers from the 29th international conference on English language research on computerized corpora (ICAME 29)* Amsterdam: Rodopi. 477-502.

Levin, B. 1993. *English Verb Classes and Alternations, A Preliminary Investigation. University of Chicago Press.*

Matsumoto, M. 2008. *From Simple Verbs to Periphrastic Expressions*. Bern and Frankfurt: Peter Lang.

Meyers, A., C. Macleod, R. Yangarber, R. Grishman, L. Barrett, and R. Reeves. 1998. Using NomLex to produce nominalization patterns for information extraction. In: *Proceedings of COLING-ACL98 Work- shop: the Computational Treatment of Nominals*.

Mukherjee, J. & St. Th. Gries. 2009. 'Collostructional nativisation in New Englishes: verb-construction associations in the International Corpus of English', *English World-Wide* 30 (1), 27-51.

Mukherjee. J & S. Hoffmann. 2009. 'Patterns across varieties: verb-complementational profiles of old and new Englishes'. Reinfandt, Chr. (ed.) *Anglistentag 2008 Tübingen: Proceedings*. Trier: WVT. 415-424.

Nagy T., I., V. Vincze and R. Farkas. 2013. Full-coverage Identification of English Light Verb Constructions. *Proceedings of IJCNLP 2013*, Oct. 14-18, Nagoya, Japan.

Nelson, G., S. Wallis, and B. Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Varieties of English Around the World: G29. John Benjamins, Amsterdam.

O'Donnell, M. B. and N. Ellis. 2010. 'Towards an inventory of English verb-argument constructions.' *Proceedings of the ACL 2010.* Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics Session. Los Angeles: ACL.

Pecina, Pavel. *Lexical Association Measures: Collocation Extraction*. Volume 4 of Studies in Computational and Theoretical Linguistics. UFAL, Praha, Czech Republic, 2009.

Ronan, P. 2012. Make Peace *and* Take Victory: *Support Verb Constructions in Old English in Comparison with Old Irish.* NOWELE Supplement Series 24. Amsterdam: Benjamins.

Ronan, P. 2014. 'Light-verb constructions in the History of English'. Davidse, K., C. Gentens, L. Ghesquière, and L. Vandelanotte (eds.) *Corpus interrogation and grammatical patterns.* Amsterdam: John Benjamins. 15-34.

Schneider, G. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Schneider, G. and L. Zipp. 2013. Discovering new verb-preposition combi- nations in New Englishes. In Joybrato Mukherjee and Magnus Huber, editors, *Studies in Variation, Contacts and Change in English, Volume 14 – Corpus Linguistics and Variation in English: Focus on non-native Englishes*. Varieng, Helsinki.

Seretan, V. 2011. *Syntax-Based Collocation Extraction*. Dordrecht: Springer.

Wierzbicka, A. 1982. 'Why Can you *Have a Drink* When You Can't **Have an Eat*?'. *Language* 58. 753-99.